Exploring a New Class of Molecules Related to Cancer, Stem Cells, and Epigenetics: Long Noncoding RNAs

Irina V. Novikova, Scott P. Hennelly, Karissa Y. Sanbonmatsu, T-6

Human long noncoding RNAs (lncRNA) are a newly discovered class of biological molecules shown to play important roles in a wide range of biological functions, including hormone signaling, embryonic stem cell differentiation, development, brain function, and cancer [1–5]. In recent years it has been shown that as much as 98% of the human genome does not code for protein. In the past three years alone, approximately 50,000 lncRNAs have been discovered. As each lncRNA is typically 1,000–10,000 residues in length (1–10 kB), lncRNAs stand to occupy a substantial fraction of the human genome. lncRNAs are often associated with chromatin remodeling and other epigenetic processes. In some cases, lncRNAs recruit the polycomb repressive complex to chromatin, facilitating modification of histone proteins. Many lincRNAs are involved in stem-cell reprogramming, either regulating or being regulated by three key stem cell proteins: Nanog, Oct4, and Sox2. One of the most famous lncRNA systems is the X chromosome. In females an entire X chromosome is almost completely shut down by a suite of gigantic lncRNAs (e.g., Xist ~17 kB, Tsix ~40 kB). Here, the Xist lncRNA is overexpressed in massive quantities and covers the entire chromosome in a “chromatin coating” process.

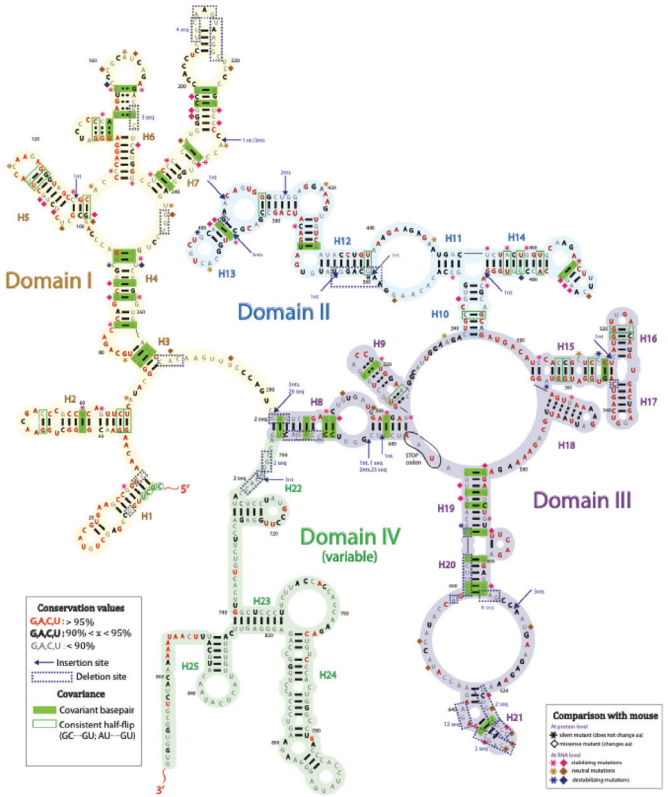

Fig. 1. Secondary structure of the steroid receptor RNA activator lncRNA determined from chemical probing and covariance analysis. Dashed purple boxes, deletions that occurred in at least 1 of 36 vertebrate sequences. If deletion occurs in more than one species, the number of species undergoing deletion at this position is specified. Purple arrows: insertion positions. Number of nucleotides, x, incorporated at insertion site is indicated by “x nts.” If insertion occurs in more than one species, the number of species undergoing insertion at this position is specified. Green-filled boxes, covariant base pairs. Green-outlined boxes, base pairs undergoing a change from a Watson-Crick base pair to a GU or UG, which do not have any instances of mismatches across all the organisms. Mutations from mouse to human tend to stabilize the RNA structure of the lncRNA.

We have produced the first secondary structure of a human long noncoding RNA ("link RNA," lncRNA) using a combination of extensive chemical probing experiments and computational sequence comparison analyses; however, the structures of these molecules remain a mystery [1-5]. As no structural studies have been performed, the following questions remain unanswered: 1) Are lncRNAs highly structured or disordered? 2) Do they contain globular sub-domains or are they organized linearly in chains of stem-loops? 3) Do lncRNAs exist in ribonucleoprotein complexes or as isolated RNAs that transiently interact with proteins? and 4) Do these molecules contain a compact core or are they more extended? We have answered the first two questions for a particular lncRNA using a combination of experimental and computational analyses [6]. In addition, very little is known about the evolution of lncRNAs. Using comparative structure analysis, we study the evolution of a lncRNA.

The subject of our study is the steroid receptor RNA activator (SRA), a lncRNA strongly associated with breast cancer [7-10]. While this molecule acts mainly as a noncoding RNA, alternative splicings of the molecule code for a protein called SRA protein, or SRAP. The steroid receptor RNA activator (874 residues) is a key player in hormone receptor regulation. This lncRNA is a nuclear co-activator and upregulates a wide variety of nuclear receptors, including the estrogen, androgen, progesterone, and thyroid hormone receptors, as well as the dosage-sensitive sex reversal (DAX-1) and steroidogenic factor (SF-1); these are key players in sex determination during development. The expression of the coding and noncoding forms of SRA occurs in different ratios for different tumor types. For this reason, many are interested in using SRA as an early onset marker for breast cancer. Understanding the molecular mechanism of SRA will help guide its usage as a marker.

Secondary, or 2D, structure in RNA is very different from 2D structure in proteins. While protein 2D structure reports on the helical versus sheet-like nature of the conformation, RNA 2D structure reports simply on Watson-Crick base pairing. However, the aggregate map of base pairing for the entire molecule yields an enormous amount of information on the architecture of the RNA, including the existence of sub-domains,

For more information contact Karissa Y. Sanbonmatsu at kys@lanl.gov

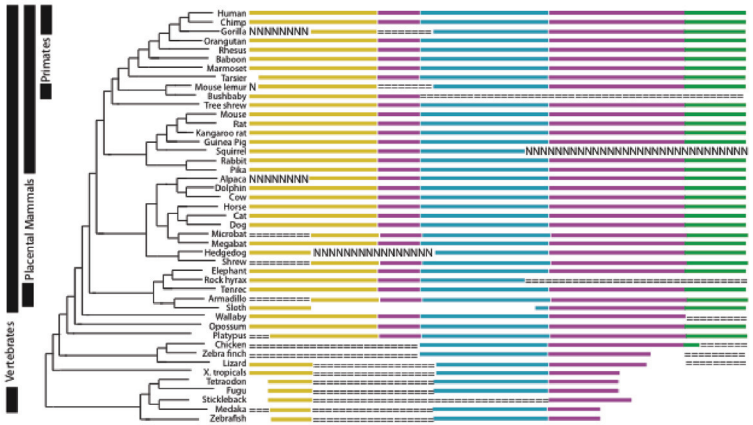

Fig. 2. Phylogenetic tree displays evolution of RNA structural domains across 45 vertebrates possessing the SRA gene. Yellow bars, domain I; cyan bars, domain II; purple/magenta bars, domain III; green bars, domain IV; "===" and "NNN" denote uncertainty in sequence alignment according to ENCODE database conventions.

the type of junction between helices, and particular bulged structural motifs. In some cases, the 2D structure alone reveals the functional mechanism itself (e.g., riboswitches). In these cases it is not necessary to solve the 3D X-ray structure to determine the mechanism of function.

The ribosome is one of the only large RNAs studied in mechanistic detail to date, so we followed in the footsteps of ribosome researchers to determine the 2D structure of SRA. We performed dimethyl sulfate chemical probing and RNase V1 digestion to locate base-paired and non-base-paired regions of the RNA. In addition, we employed newer probing techniques developed to study riboswitches. In-line probing and selective 2' hydroxyl acylation by primer extension both produce single-nucleotide resolution; they report on the RNA backbone mobility, enabling us to identify base-paired residues. We used multiple sequence alignment across 45 species to verify our helices through covariance. So, if a GC base pair in humans changes to another Watson-Crick base pair (AU, UA, or CG) in another species, this is evidence that supports our structure.

Our results show the lncRNA to be highly structured (Fig. 1) and organized into four sub-domains. In all, we identified 25 RNA helices, 16 terminal loops, 15 internal loops, and 5 junction regions. Our structure is consistent with previous in vivo site-directed mutagenesis and deletion studies. Our structure has several functional implications. It is currently not known whether SRA co-activates hormone receptors by escorting them through the nuclear membrane or by acting as a structural scaffold for chromatin in the transcription complex. We have shown that SRA is highly structured, suggesting it is possible that SRA provides structural scaffolding for the transcription complex. In addition, it was previously suggested that H13 (or Str 7) is essential for hormone-receptor binding. However, a variety of other proteins interact with SRA, including DAX-1, SF-1, deadbox proteins P68/P72, Myo-D, SLIRP, SHARP, Pus1, and Pus3. We find the three-way junction branching helices H15, H16, and H17 to be much more highly conserved than H13. Thus, this junction region may be a second protein-binding site.

Another interesting aspect of the SRA gene is that it functions as an RNA in some cases but as a protein in others. It has been proposed that at some point in evolution, the noncoding Xist lncRNA (responsible for X-chromosome inactivation) might have originated gradually, allowing for a period of time where noncoding and coding isoforms of the gene coexisted. The coding and noncoding isoforms of SRA also originated from the same gene, suggesting that this lncRNA might be a rare and unique capture of this stage of evolution. To examine this more carefully, we studied the evolution of SRA (Fig. 2). Since the 3D structure of the mouse SRAP has been solved by nuclear magnetic resonance, we had the unique opportunity to study the co-evolution of RNA structure and protein structure. We found rapid evolutionary stabilization of the RNA structure. The vast majority of mutations from mouse to human act to stabilize RNA helices. In contrast, most mutations of protein sequence occur in linking regions between helices and do not appear to be obviously stabilizing. In addition, several frame-disrupting mutations occur from mouse to human in conserved regions, suggesting that evolutionary pressure preserves the RNA structural core rather than its translational product.

[1] Ulitsky, I. et al., *Cell* **147,** 1537 (2011).
[2] Guttman, M. et al., *Nature* **477,** 295 (2011).
[3] Huarte, M. et al., *Cell* 1**42,** 409 (2010).
[4] Gupta, R.A. et al., *Nature* **464**, 1071 (2010).
[5] Ponting, C.P. et al., *Cell* **136,** 629 (2009).
[6] Novikova, I.V. et al., *Nucleic Acids Res,* in press (2012).
[7] Lanz, R.B. et al., *Cell* **97,** 17 (1999).
[8] Chooniedass-Kothari, S. et al., *FEBS Lett* **566,** 43 (2004).
[9]Hube, F. et al., *DNA Cell Biol* **25,** 418 (2006).
[10] Yan, Y. et al., *Breast Cancer Res* **11,** R67 (2009).

Funding Acknowledgments
LANL Laboratory Directed Research and Development Program